

DOI:10.13288/j.11-2166/r.2024.17.007

# 基于 Stacking 集成算法的中医证候诊断模型建立 ——以肺癌为例

郭小川<sup>1,2,3</sup>, 冯贞贞<sup>1,2</sup>, 刘文瑞<sup>1,2,3</sup>, 李建生<sup>1,2</sup>✉

1. 河南中医药大学/呼吸疾病中医药防治省部共建协同创新中心, 河南省郑州市郑东新区金水东路156号, 450046; 2. 河南中医药大学第一附属医院; 3. 河南中医药大学第一临床医学院

**[摘要]** **目的** 探索 Stacking 集成算法优化中医证候诊断模型效能的方法。**方法** 以肺癌中医证候诊断模型的构建为例, 将来自 9 家医院肺癌患者的 2598 例次临床症状及体征信息作为自变量 (即特征变量), 中医证候信息作为因变量, 采用 Python 3.7 软件将临床数据以 8:2 比例按照随机数字表法分为训练集和测试集。运用卡方检验、Spearman 相关性检验、最小绝对值收缩和选择算子 (LASSO) 逻辑回归分析筛选肺癌中医证候的稳定特征; 利用支持向量机 (SVM)、K 近邻算法 (KNN)、随机森林 (RF)、极端随机树 (ExtraTrees)、极端梯度提升机 (XGBoost)、轻量级梯度提升机 (LightGBM)、自适应增强 (AdaBoost)、梯度提升 (GB) 及多层神经网络 (MLP) 9 种机器学习算法进行训练, 得到 9 种基础模型。在上述基础模型中筛选出性能表现较优的 4 种模型, 运用 Stacking 集成算法进行融合形成融合模型, 并通过上述 9 种机器学习算法对融合模型进行二次训练, 运用准确率、微平均受试者工作特征 (micro-average ROC) 曲线、曲线下面积 (AUC) 和混淆矩阵指标进行评估, 筛选最优诊断模型。**结果** 经数据处理得到稳定特征 79 个、中医证候 13 个。在基础模型训练中, RF、ExtraTrees、MLP 及 SVM 基础模型综合性能表现较优, 故将该 4 种模型的证候预测分布作为二次训练数据, 并基于 Stacking 集成算法得到 9 种融合模型 (SVM, KNN, RF, ExtraTree, XGBoost, LightGBM, GB, AdaBoost, MLP)。其中 XGBoost 融合模型性能表现最优, 在训练集和测试集中准确率分别为 0.850 和 0.838, 过拟合差异为 0.012, micro-average ROC 曲线下面积 (micro-average AUC) 为 0.996。所有融合模型在测试集中的准确率和 micro-average AUC 较基础模型均有改善。**结论** 以肺癌的中医证候数据为例, 通过 Stacking 集成算法得出 XGBoost 融合模型在提升肺癌中医证候诊断效能方面具有显著优势。可见 Stacking 集成算法能整合多种模型算法的优点, 有效提升中医证候诊断模型识别效能, 为同类研究提供方法学借鉴。

**[关键词]** 中医证候诊断模型; 肺癌; 证候; 机器学习; Stacking 集成算法

辨证施治是中医诊疗疾病的主要模式, 证候的辨识为中医诊疗关键环节, 在中医证候诊断中, 尚存在辨证标准不一、主观性强、证候繁杂、临床症状与辨证分型缺乏量化关联等问题<sup>[1]</sup>。现代计算机技术的进步和人工智能算法的普及, 为解决以上问题提供了新思路<sup>[2-3]</sup>。目前, 支持向量机、K 近邻、随机森林等机器学习算法在中医证候内部联系的挖

掘中已得到广泛应用, 但单一模型存在精度不佳、过拟合风险较高等问题<sup>[4]</sup>。Stacking 集成算法通过融合多种机器学习模型的优势, 不仅能显著提高预测的准确性, 也能降低模型对训练集过拟合的风险, 克服了依赖单一算法的局限性<sup>[5]</sup>。本研究以肺癌为例, 基于前期收集的临床数据, 充分考虑各类模型在数据模式和特征中可能存在的敏感性差异, 最终选择分类、回归和通用模型大类中具有代表性的 9 种模型<sup>[6]</sup>, 结合 Stacking 集成算法, 初步建立并优化中医证候诊断模型, 为同类研究提供方法学参考。

**基金项目:** 国家中医药管理局中医药传承与创新“百千万”人才工程-岐黄工程首席科学家 (国中医药人教函[2020]219号); 国家自然科学基金(82205313); 河南省中医药科学研究专项课题(2022JDZX102)

✉ 通讯作者: li\_js8@163.com

## 1 资料与方法

### 1.1 诊断标准

1.1.1 西医诊断标准 原发性肺癌诊断标准参照《中华医学会肺癌临床诊疗指南（2019 版）》<sup>[7]</sup>：

经活检或细胞学病理检查，存在肺癌细胞者即可确诊。

1.1.2 中医辨证标准 依据《中医临床诊疗术语第 2 部分：证候》<sup>[8]</sup> 的肺癌辨证分型标准，结合实际临床调查，共匹配 18 种中医证候类型。并根据专家咨询结果，剔除比例小于 1% 的证型，最终得到 13 种中医证型：气滞血瘀证、痰瘀互结证、痰热壅肺证、痰湿阻肺证、肺热炽盛证、阴虚内热证、肺气虚证、肺阴虚证、气阴两虚证、肺脾气虚证、肺肾气虚证、肺肾阴虚证、气血两虚证。

### 1.2 纳入标准

1) 符合上述诊断标准及中医辨证标准；2) 年龄  $\geq 18$  岁；3) 能够用汉语进行交流与阅读；4) 自愿参加本调查并签署知情同意书。

### 1.3 排除标准

1) 身心状态不适合进行研究，如妊娠期、哺乳期女性，病情危重、终末期患者，重度抑郁障碍、焦虑障碍患者，以及存在认知或其他损害（如视力、听力等）而影响调查的患者等；2) 合并活动性肺结核、支气管哮喘、支气管扩张症、肺脓肿、尘肺等呼吸系统疾病的患者。

### 1.4 数据来源

2020 年 6 月 30 日至 2021 年 11 月 30 日在 9 家医院开展横断面临床调查，发放调查问卷 2200 份，回收 2150 份，有效问卷 2058 份。问卷调查内容包括患者基本情况、中医四诊信息和辨证分型 3 个方面。每位肺癌患者的中医证候均由 2 位工作年限  $\geq 3$  年且具有中级及以上职称的中医师联合判断。当同一患者存在多个证候时，除主证外的每个兼证分别作为独立样本纳入。整理后得到有效患者 2598 例次（河南省中医院 672 例次、陕西中医药大学附属医院 354 例次、河南中医药大学第一附属医院 292 例次、山东中医药大学附属医院 285 例次、河南省肿瘤医院 253 例次、上海中医药大学附属龙华医院 252 例次、广西中医药大学第一附属医院 239 例次、辽宁中医药大学附属医院 134 例次、中国中医科学院广安门医院 117 例次），患者信息见附图 1 及附表 1（请扫描本文二维码获取）。提取纳入中医证型 13 种，临床症状及体征 125 个（包括舌象

23 个、脉象 19 个）。

### 1.5 数据库建立及数据预处理

使用 Epidata 3.1 软件对所收集的临床问卷信息（患者症状及体征、证型）进行录入，以此建立原始数据库，后将数据库信息以 UTF-8 编码形式导出至 CSV 数据集文件中，进行数据分析。

在保证各证候数量与原数据集比例相同条件下，使用 Python 3.7 软件将数据集以 8 : 2 的比例按照随机数字表法分为训练集和测试集<sup>[9]</sup>，将患者 125 个临床症状及体征作为自变量（即特征变量）。数据集中采用“0-1”规则对特征变量进行编码（样本有该特征表示为“1”，无该特征表示为“0”）。对中医证型进行数字编码并作为因变量以便数据分析，编码形式如下：1-气滞血瘀证、2-痰瘀互结证、3-痰热壅肺证、4-痰湿阻肺证、5-气血两虚证、6-肺热炽盛证、7-阴虚内热证、8-肺气虚证、9-肺阴虚证、10-气阴两虚证、11-肺脾气虚证、12-肺肾气虚证、13-肺肾阴虚证。当单个特征存在样本缺失，若缺失比例  $< 10\%$ ，为定性特征的采用众数填补，为定量特征的采用均数填补；若缺失比例  $\geq 10\%$ ，则删除该特征。采用卡方检验进行显著性分析， $P > 0.05$  则认为该特征与证候间无显著差异，剔除该特征，反之则保留。

首先，基于 Python 3.7 软件的 SciPy 包对纳入特征进行 Spearman 相关性检验，当两两特征间相关系数  $> 0.9$  时，认为特征间存在共线性风险，仅保留其中频次较高的特征<sup>[10]</sup>。其次，使用 Python 3.7 软件的 Keras 包对纳入特征进行最小绝对值收缩和选择算子（LASSO）逻辑回归分析，设置惩罚系数  $\alpha = 1.0$ ，计算模型复杂度与模型泛化能力间达最优平衡时的 Lambda 值，并计算该 Lambda 值下各特征的权重（ $w$ ），当  $w \neq 0$  时，被判定为稳定特征而予以保留，反之则剔除。相关数据预处理公式见：式（1）、式（2）、式（3）（见附件 1，请扫描本文二维码获取）。

1.5.1 数据标准化<sup>[11]</sup> 数据标准化将数据转化为特定属性的标准形式，通过对原始数据的处理，使其转化为“均值为 0，标准差为 1”的数据集，便于数据处理和建模。

1.5.2 Spearman 相关性检验<sup>[12]</sup> 采用 Spearman 相关性检验识别并剔除数据中相关系数  $> 0.9$  的高相关性特征，降低过拟合风险，进而提升所训练模型的泛化能力。

1.5.3 LASSO 逻辑回归<sup>[13]</sup> LASSO 逻辑回归是一种优化后的线性回归模型,通过对特征权重施加 L1 正则化实现稳定特征的筛选,从而降低模型训练复杂度,提高模型泛化能力和预测效能。

## 1.6 模型训练与评估

1.6.1 模型开发环境与框架 本研究基于 Anaconda 4.10.3 软件,采用 PyTorch (2.1.2+GPU 版本) 算法框架,对基础模型和 Stacking 集成模型进行训练。

1.6.2 机器学习算法 当下机器学习领域,多种算法被广泛应用于各类数据的训练预测中,这些算法在解决分类、回归及模式识别等问题中展现了不同优势。中医数据蕴含着丰富的中医理论和实践知识,涉及症状、舌象、脉象等多个数据维度,各维度均可能对辨证诊断产生未知程度的影响,具有非线性和复杂性<sup>[14]</sup>。选用多种模型共同挖掘患者临床症状、体征与中医证候间的联系,有助于发挥不同模型各自的优势,提高中医辨证的准确性。本研究涉及分类、回归和通用类别的模型共 9 种,包括支持向量机 (SVM)、K 近邻算法 (KNN)、随机森林 (RF)、极端随机树 (ExtraTrees)、极端梯度提升 (XGBoost)、轻量级梯度提升机 (LightGBM)、自适应增强 (AdaBoost)、梯度提升 (GB) 及多层神经网络 (MLP)。这些模型中,SVM 和 KNN 是广泛应用于分类和回归问题的监督学习模型<sup>[15]</sup>,其中 SVM 通过在特征空间中寻找最优超平面,最大化类别之间的边界间隔<sup>[16]</sup>;KNN 可测量不同特征值之间的距离实现对因变量的预测<sup>[17]</sup>。RF 和 ExtraTrees 均属于决策树模型,RF 通过在每次树分裂时从特征的随机子集中选择最优特征<sup>[18]</sup>;ExtraTrees 在每个决策节点上对特征的分裂点进行随机选择,进一步增加模型的随机性<sup>[19]</sup>。GB、LightGBM 和 XGBoost 是基于梯度提升的方法训练模型,GB 通过逐步添加弱学习器(如决策树)来修正前一轮迭代的残差,逐步减小预测误差损失(如均方误差或交叉熵损失)<sup>[20]</sup>;LightGBM 使用直方图的树分裂算法和叶子优先策略,可以更快减少预测误差损失<sup>[21]</sup>;XGBoost 则引入了正则化项(L1 和 L2)以控制模型复杂度,提高训练模型效率和性能<sup>[22]</sup>。AdaBoost 属集成学习算法,通过组合多个弱学习器提升模型性能,自动调整错误分类的关注度<sup>[23]</sup>;MLP 是深度学习的核心算法,通过非线性激活函数建立多层结构,进而实现预测的目的<sup>[24]</sup>。

1.6.3 模型评价指标 使用模型评价指标包括准确率、微平均受试者工作特征 (micro-average ROC) 曲线、ROC 曲线下面积 (AUC) 和混淆矩阵来筛选较优模型。

在模型评价指标中,真正例 (TP) 是模型将正例数据预测为正例的数量;假正例 (FP) 是模型将负例预测为正例的数量;真负例 (TN) 是模型将负例数据预测为负例的数量;假负例 (FN) 是模型将正例数据预测为负例的数量。准确率是模型性能的基本指标,用于评估模型在整体数据集上的预测效能<sup>[25]</sup>。micro-average ROC 是一种在数据类别不平衡的多分类问题中,用来综合评估分类模型性能的方法<sup>[26]</sup>。AUC 是反映分类模型综合性能的指标,可评估模型的预测准确性,而微平均 ROC 曲线下面积 (micro-average AUC) 作为一种在数据分布不平衡条件下的评估指标,可对模型在各类样本中的预测能力进行更均衡的量化评估<sup>[27]</sup>。混淆矩阵是一种分类模型性能评估手段,能反映对模型在不同类别中预测结果的具体信息<sup>[28]</sup>。各评判指标解释及计算见式 (4)、式 (5)、式 (6)、式 (7) (见附件 1,请扫描本文二维码获取)。

1.6.4 基础模型训练 在训练集中,采用 10 折 ( $K=10$ ) 交叉验证的方法对 SVM、KNN、RF、ExtraTrees、XGBoost、LightGBM、AdaBoost、GB、MLP 共 9 种模型进行训练和评估。测试集用于评估模型性能,可通过比较不同模型在测试集的表现以检验过拟合的风险。模型性能通过准确率、micro-average ROC 曲线、micro-average AUC 和混淆矩阵进行综合评估。基于评估结果,最终筛选出 4 种表现较优的模型作为后续 Stacking 集成学习的基础模型。该筛选机制优势为在 Stacking 集成模型构建过程中,仅融合最具代表性和较优效能的基础模型,从而既能保证模型多样性,又能最大限度提高最终模型预测的准确性和泛化能力。

1.6.5 Stacking 集成模型训练 Stacking 集成算法是一种能融合多种模型的先进算法,其所包含的不同算法能从多个角度挖掘证候数据间的联系和规律,从而实现更精准的辨证。Stacking 集成算法通过将基础模型的预测结果叠加为新数据,以此训练新模型,最终提高数据整体预测判断的能力,该算法对高维度的复杂性数据尤其适用<sup>[29]</sup>。本研究基于该算法对 4 种表现较优的基础模型进行融合,使用 4 种较优基础模型的预测结果堆叠成新的数据

集, 对 SVM、KNN、RF、ExtraTrees、XGBoost、LightGBM、AdaBoost、GB 和 MLP 模型进行二次训练。最终根据模型准确率、micro-average ROC 曲线、micro-average AUC 和混淆矩阵指标, 输出综合表现最优的模型。Stacking 算法原理和模型构建流程见附图 2 (请扫描本文二维码获取)。

## 2 结果

### 2.1 临床症状、体征及证候数据特征

2598 例次肺癌患者中肺脾气虚证 581 例次 (22.4%)、气阴两虚证 448 例次 (17.2%)、痰瘀互结证 342 例次 (13.2%)、气滞血瘀证 314 例次 (12.1%)、痰湿阻肺证 238 例次 (9.2%)、肺气虚证 188 例次 (7.2%)、痰热壅肺证 142 例次 (5.5%)、阴虚内热证 95 例次 (3.7%)、肺肾气虚证 60 例次 (2.3%)、肺阴虚证 55 例次 (2.1%)、肺热炽盛证 53 例次 (2.0%)、肺肾阴虚证 42 例次 (1.6%)、气血两虚证 40 例次 (1.5%)。常见临床症状为咳嗽、乏力、神疲、咳痰等; 舌象为舌质红、舌质暗红、舌苔白、舌苔淡白、舌苔薄等; 脉象为脉细、脉弦、脉沉细、脉弦滑等, 见附图 3 (请扫描本文二维码获取)。

### 2.2 LASSO 逻辑回归分析与特征筛选结果

咯血、喘气、神昏、舌苔白腻、脉浮与证候之间的卡方检验显示, 差异无统计学意义 ( $P>0.05$ ), 予以剔除。Spearman 相关性检验未发现需剔除的高相关性特征。LASSO 逻辑回归分析结果显示, 随 Lambda 值的增加 (正则化程度加强), 各特征系数逐渐减小, 当 Lambda = 0.069 时, 模型复杂度与模型泛化能力间达到最优平衡 (见附图 4, 请扫描本文二维码获取); 且当 Lambda = 0.069 时, 以特征  $w = 0$  为剔除标准, 剔除不稳定特征 41 个,  $w \neq 0$  时保留稳定特征 79 个, 稳定特征中舌苔淡白、舌苔薄、多梦、盗汗、痰量少、脉细等特征权重较高。见表 1。

### 2.3 基础模型训练结果

表 2 示, 经训练和测试后所得 9 种基础模型中, 训练集准确率较高的 4 种模型分别为 ExtraTrees (0.784)、RF (0.779)、MLP (0.634) 和 SVM (0.600); 测试集准确率较高的 4 种模型分别为 RF (0.516)、ExtraTrees (0.471)、SVM (0.399) 和 MLP (0.379)。表 3 示, micro-average AUC 较高的 4 种模型分别为 SVM (0.870)、RF (0.859)、XGBoost

表 1 经筛选的 2598 例次肺癌临床症状及体征的 LASSO 逻辑回归分析的特征权重表 (Lambda = 0.069)

Table 1 LASSO feature weight table of 2598 screened cases of clinical symptoms and signs of lung cancer (Lambda = 0.069)

| 症状及体征 | $w$   | 症状及体征 | $w$    |
|-------|-------|-------|--------|
| 自汗    | 0.022 | 脉弦    | -0.367 |
| 小便不利  | 0.026 | 舌瘀斑   | -0.323 |
| 痰色白   | 0.029 | 胸胁痛   | -0.260 |
| 便溏    | 0.029 | 舌苔薄黄  | -0.235 |
| 恶心    | 0.034 | 舌质紫暗  | -0.231 |
| 呕吐    | 0.034 | 壮热    | -0.185 |
| 气短    | 0.040 | 脉涩    | -0.183 |
| 口苦    | 0.044 | 面色晦暗  | -0.177 |
| 面色萎黄  | 0.054 | 痛有定处  | -0.166 |
| 舌体齿痕  | 0.055 | 舌苔厚   | -0.151 |
| 舌体瘦小  | 0.059 | 胃脘痞满  | -0.139 |
| 脉濡    | 0.064 | 舌苔腻   | -0.130 |
| 舌苔花剥  | 0.067 | 喉中痰鸣  | -0.129 |
| 四肢麻木  | 0.068 | 渴不欲饮  | -0.108 |
| 腰酸    | 0.075 | 大便秘结  | -0.104 |
| 脉细数   | 0.076 | 小便频数  | -0.097 |
| 舌质裂纹  | 0.077 | 咳嗽    | -0.097 |
| 口干    | 0.080 | 发热    | -0.083 |
| 颜面浮肿  | 0.086 | 脉细涩   | -0.081 |
| 消瘦    | 0.090 | 呃逆    | -0.078 |
| 头痛    | 0.092 | 咳痰    | -0.077 |
| 脉沉细   | 0.094 | 脉数    | -0.070 |
| 神疲    | 0.109 | 舌质暗红  | -0.070 |
| 舌苔白   | 0.110 | 脉滑数   | -0.054 |
| 耳鸣    | 0.120 | 肢体肿胀  | -0.052 |
| 舌苔少   | 0.123 | 肌肤甲错  | -0.050 |
| 腰痛    | 0.135 | 痰腥臭   | -0.042 |
| 心悸    | 0.136 | 痰质浓稠  | -0.041 |
| 乏力    | 0.142 | 咽干    | -0.040 |
| 纳差    | 0.144 | 四肢逆冷  | -0.035 |
| 脉细弱   | 0.152 | 脉濡缓   | -0.029 |
| 咳声低微  | 0.160 | 脉弦滑   | -0.025 |
| 脉细    | 0.182 | 气喘    | -0.024 |
| 痰量少   | 0.184 | 面色红   | -0.023 |
| 盗汗    | 0.188 | 舌苔薄腻  | -0.020 |
| 舌苔薄   | 0.191 | 脉滑    | -0.012 |
| 多梦    | 0.210 | 口渴    | -0.010 |
| 舌苔淡白  | 0.226 | 痰黏难咳  | -0.006 |
|       |       | 小便清长  | -0.005 |
|       |       | 舌苔黄腻  | -0.001 |
|       |       | 舌质红绛  | -0.001 |

注: LASSO, 最小绝对收缩和选择算子, 用于特征选择和降低过拟合风险;  $w$ , 特征权重, 反映特征对肺癌证候诊断的重要性,  $w>0$  表示该特征与证候诊断呈正相关,  $w<0$  表示该特征与证候诊断呈负相关。

(0.847) 和 LightGBM (0.847)。

经评估，虽然 XGBoost 和 LightGBM 在 micro-average AUC 上表现较优，但二者准确率在训练集中分别为 0.565 和 0.538，测试集中分别为 0.359 和 0.353，而 RF、ExtraTrees、MLP 和 SVM 在准确率和 micro-average AUC 上表现较 XGBoost、LightGBM 更平衡，泛化能力更强，故将 RF、ExtraTrees、

表 2 基于 2598 例次肺癌临床症状及体征的 9 种基础模型与融合模型的训练集、测试集准确率比较

Table 2 Comparison of the accuracy of the training and test sets of the nine basic models and the fusion model based on 2598 cases of clinical symptoms and signs of lung cancer

| 模型         | 基础模型  |       | 融合模型  |       |
|------------|-------|-------|-------|-------|
|            | 训练集   | 测试集   | 训练集   | 测试集   |
| SVM        | 0.600 | 0.399 | 0.814 | 0.803 |
| KNN        | 0.573 | 0.373 | 0.835 | 0.855 |
| RF         | 0.779 | 0.516 | 0.854 | 0.840 |
| ExtraTrees | 0.784 | 0.471 | 0.855 | 0.828 |
| XGBoost    | 0.565 | 0.359 | 0.850 | 0.838 |
| LightGBM   | 0.538 | 0.353 | 0.831 | 0.794 |
| GB         | 0.381 | 0.268 | 0.821 | 0.819 |
| AdaBoost   | 0.280 | 0.222 | 0.460 | 0.466 |
| MLP        | 0.634 | 0.379 | 0.812 | 0.807 |

注：SVM，支持向量机；KNN，K 近邻算法；RF，随机森林；ExtraTrees，极端随机树；XGBoost，极端梯度提升；LightGBM，轻量级梯度提升机；AdaBoost，自适应增强；GB，梯度提升；MLP，多层神经网络。

MLP 和 SVM 作为 Sstacking 集成算法的 4 种基础模型进行融合模型训练。

2.4 基于 Stacking 集成算法的融合模型训练结果

表 2 示，基于 Stacking 集成算法的融合模型训练和测试结果显示，除 AdaBoost 和 LightGBM 外，融合模型在训练集和测试集中准确率均超过 0.800；训练集中准确率较高的 3 种融合模型为 ExtraTrees (0.855)、RF (0.854) 和 XGBoost (0.850)，测试集中准确率较高的 3 种融合模型为 KNN (0.855)、RF (0.840) 和 XGBoost (0.838)。其中，XGBoost 过拟合差异为 0.012，RF 为 0.014，ExtraTrees 为 0.027；而 KNN 为 -0.002，提示该模型存在欠拟合风险。

表 4 示，除 AdaBoost 外，融合模型的 micro-average AUC 均超过 0.900，micro-average AUC 较高的 3 种融合模型为 XGBoost (0.996)、MLP (0.995) 和 GB (0.994)。9 种融合模型中，准确率和 micro-average AUC 均位列前三位的是 XGBoost，其各中医证候的混淆矩阵见图 1。

2.5 融合模型与基础模型效能对比

研究结果显示，融合模型较基础模型对训练集准确率的改善率范围为 9.91%~115.49%，平均改善率 45.87%；对测试集准确率的改善率范围为 62.79%~205.60%，平均改善率 117.32%；测试集 micro-average AUC 改善率范围为 1.50%~32.71%，平均改善率 16.68%，见图 2。

表 3 9 种基础模型对肺癌 13 种中医证候分类的各证候 AUC 与 micro-average AUC 比较

Table 3 Comparison of syndrome AUC and micro-average AUC for the classification of 13 syndromes in lung cancer by 9 basic models

| 项目                | SVM   | KNN   | RF    | ExtraTrees | XGBoost | LightGBM | AdaBoost | GB    | MLP   |
|-------------------|-------|-------|-------|------------|---------|----------|----------|-------|-------|
| 气滞血瘀证             | 0.786 | 0.739 | 0.759 | 0.753      | 0.732   | 0.696    | 0.572    | 0.688 | 0.774 |
| 痰瘀互结证             | 0.751 | 0.694 | 0.816 | 0.708      | 0.637   | 0.653    | 0.528    | 0.535 | 0.728 |
| 痰热壅肺证             | 0.812 | 0.699 | 0.894 | 0.844      | 0.837   | 0.913    | 0.617    | 0.839 | 0.747 |
| 痰湿阻肺证             | 0.883 | 0.874 | 0.866 | 0.818      | 0.794   | 0.807    | 0.645    | 0.626 | 0.812 |
| 肺热炽盛证             | 0.993 | 0.990 | 0.993 | 0.993      | 0.990   | 1.000    | 0.940    | 0.957 | 0.977 |
| 阴虚内热证             | 0.971 | 0.984 | 1.000 | 1.000      | 0.911   | 0.993    | 0.729    | 0.760 | 0.941 |
| 肺气虚证              | 0.772 | 0.799 | 0.875 | 0.800      | 0.948   | 0.889    | 0.698    | 0.705 | 0.779 |
| 肺阴虚证              | 0.771 | 0.736 | 0.800 | 0.776      | 0.718   | 0.704    | 0.518    | 0.584 | 0.755 |
| 气阴两虚证             | 0.247 | 0.460 | 0.602 | 0.470      | 0.369   | 0.411    | 0.542    | 0.407 | 0.449 |
| 肺脾气虚证             | 0.829 | 0.761 | 0.834 | 0.861      | 0.796   | 0.742    | 0.542    | 0.709 | 0.839 |
| 肺肾气虚证             | 0.772 | 0.779 | 0.846 | 0.854      | 0.722   | 0.691    | 0.535    | 0.623 | 0.801 |
| 肺肾阴虚证             | 0.992 | 0.743 | 0.725 | 0.862      | 0.930   | 0.953    | 0.454    | 0.815 | 0.814 |
| 气血两虚证             | 0.990 | 0.993 | 0.993 | 0.483      | 0.980   | 0.993    | 0.877    | 0.997 | 0.772 |
| micro-average AUC | 0.870 | 0.812 | 0.859 | 0.830      | 0.847   | 0.847    | 0.749    | 0.798 | 0.843 |

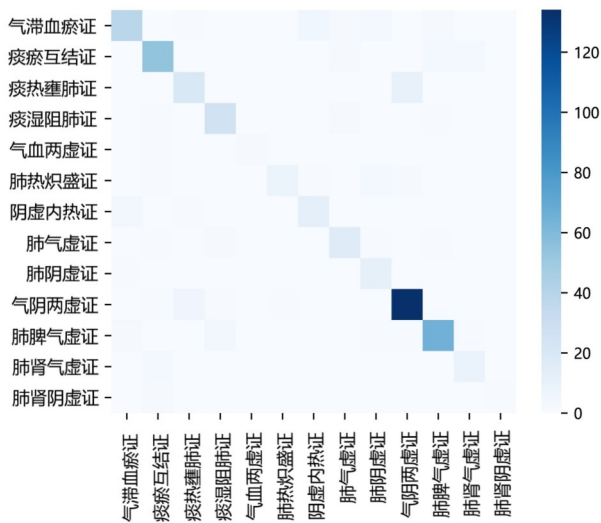
注：AUC，受试者工作特征（ROC）曲线下面积；micro-average AUC，微平均 ROC 曲线下面积，用于综合评估多类别分类模型整体性能，即评估模型对 13 种中医证候整体的预测准确性；SVM，支持向量机；KNN，K 近邻算法；RF，随机森林；ExtraTrees，极端随机树；XGBoost，极端梯度提升；LightGBM，轻量级梯度提升机；AdaBoost，自适应增强；GB，梯度提升；MLP，多层神经网络。

表 4 9 种融合模型对肺癌 13 种中医证候分类的各证候 AUC 与 micro-average AUC 比较

Table 4 Comparison of syndrome AUC and micro-average AUC for the classification of 13 syndromes in lung cancer by 9 fusion models

| 项目                | SVM   | KNN   | RF    | ExtraTrees | XGBoost | LightGBM | AdaBoost | GB    | MLP   |
|-------------------|-------|-------|-------|------------|---------|----------|----------|-------|-------|
| 气滞血瘀证             | 0.990 | 0.983 | 0.994 | 0.967      | 0.991   | 0.986    | 0.309    | 0.985 | 0.985 |
| 痰瘀互结证             | 0.987 | 0.978 | 0.986 | 0.969      | 0.994   | 0.991    | 0.344    | 0.991 | 0.991 |
| 痰热壅肺证             | 0.988 | 0.987 | 0.993 | 0.990      | 0.991   | 0.989    | 0.582    | 0.992 | 0.989 |
| 痰湿阻肺证             | 0.990 | 0.977 | 0.995 | 0.977      | 0.997   | 0.994    | 0.309    | 0.996 | 0.995 |
| 肺热炽盛证             | 1.000 | 0.833 | 0.999 | 0.999      | 0.999   | 0.999    | 0.997    | 0.999 | 0.999 |
| 阴虚内热证             | 0.988 | 0.957 | 0.962 | 0.961      | 0.994   | 0.994    | 0.428    | 0.992 | 0.991 |
| 肺气虚证              | 0.972 | 0.691 | 0.995 | 0.969      | 0.993   | 0.973    | 0.672    | 0.976 | 0.983 |
| 肺阴虚证              | 0.991 | 0.902 | 0.967 | 0.964      | 0.991   | 0.988    | 0.687    | 0.990 | 0.989 |
| 气阴两虚证             | 0.985 | 0.691 | 0.997 | 0.997      | 0.997   | 0.996    | 0.675    | 0.997 | 0.991 |
| 肺脾气虚证             | 0.995 | 0.951 | 0.996 | 0.988      | 0.996   | 0.995    | 0.969    | 0.996 | 0.996 |
| 肺肾气虚证             | 0.993 | 0.932 | 0.996 | 0.995      | 0.997   | 0.994    | 0.687    | 0.996 | 0.995 |
| 肺肾阴虚证             | 0.991 | 0.749 | 0.996 | 0.996      | 0.997   | 0.997    | 0.675    | 0.997 | 0.996 |
| 气血两虚证             | 0.999 | 0.667 | 0.998 | 0.832      | 0.999   | 0.999    | 0.999    | 0.999 | 0.999 |
| micro-average AUC | 0.993 | 0.934 | 0.993 | 0.982      | 0.996   | 0.993    | 0.810    | 0.994 | 0.995 |

注：AUC，受试者工作特征（ROC）曲线下面积；micro-average AUC，微平均 ROC 曲线下面积，用于综合评估多类别分类模型整体性能，即评估模型对 13 种中医证候整体的预测准确性。SVM，支持向量机；KNN，K 近邻算法；RF，随机森林；ExtraTrees，极端随机树；XGBoost，极端梯度提升；LightGBM，轻量级梯度提升机；AdaBoost，自适应增强；GB，梯度提升；MLP，多层神经网络。



注：横坐标和纵坐标分别代表预测类别和真实类别，图例单元格内的颜色表示对应类别的样本例次数。

图 1 基于 2598 例次肺癌临床症状及体征的 XGBoost 融合模型 13 种中医证候混淆矩阵图

Figure 1 Confusion matrix of the XGBoost fusion model for multiple syndromes in symptoms and signs of 2598 cases of lung cancer

### 3 讨论

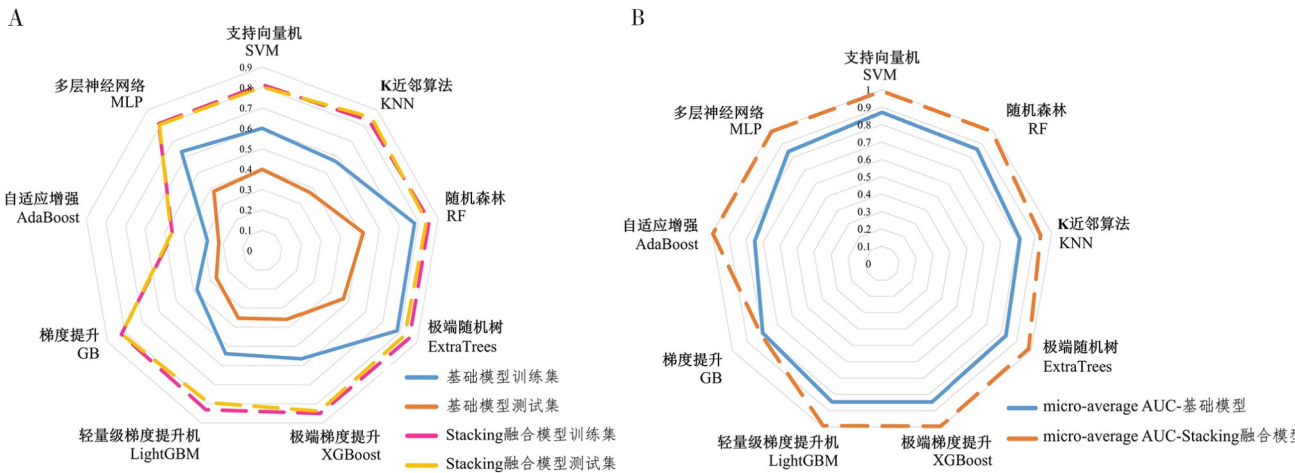
#### 3.1 传统中医证候诊断模型的局限性

精准辨证是突显中医疗效优势的前提，证候是精准辨证的核心<sup>[30-31]</sup>。在中医理论体系中，疾病病因较多，证候复杂。以肺癌为例，病理因素包括但不限于气滞、血瘀、痰浊、毒邪、脏器虚损等，这些致病因素可以单独存在，也可以相互交织构成多

种证候<sup>[32]</sup>。不同证候表现出不同的核心症状，这些症状可能是单一的也可能是多样的，当兼有多个症状时，往往有潜在关联<sup>[33]</sup>。现有证候诊断模型多基于贝叶斯算法，该算法以局部条件独立性假设为前提，即要求自变量只能与父节点自变量存在关联<sup>[34]</sup>，但对证候数据集而言，核心症状间的联系是普遍、必要且无可避免的，可能导致训练模型的泛化能力降低，加剧过拟合风险。此外，证候诊断模型所包含的自变量（症状体征）较多，传统模型存在“维度灾难”风险，导致训练所需样本量激增，建模难度增大。在传统单一模型中，不同算法虽各有专长，但也有缺陷，如 SVM 和 KNN 对不平衡数据的处理结果较差；而 RF 和 Extratrees 引入的随机性加剧了模型偏差风险；XGBoost、LightGBM 和 MLP 在样本量不足时过拟合风险较大；GB 和 AdaBoost 则在数据噪声较大时表现不佳。

#### 3.2 Stacking 集成算法为中医证候诊断模型优化提供思路

中医的证候诊断非线性强，具有复杂性、不平衡性、兼夹性等特点<sup>[14, 35]</sup>，因此需要一种与特定数据相适应的训练思路。模型融合技术作为新兴算法，通过合理调配基础模型的预测结果，获得更准确的预测，改善模型综合效能。Stacking 集成算法作为一种模型融合手段，基于人工智能特征抽象和参数迭代思想，将基础模型预测结果抽象为新的特征进行训练，捕捉证候数据多样性，提高模型泛化



注：micro-average AUC，微平均受试者工作特征曲线下面积；A，融合模型与基础模型在训练集和测试集中的准确率比较（轴上的数值代表各模型对应的准确率）；B，融合模型与基础模型在测试集中 micro-average AUC 比较（轴上的数值代表各模型对应的 micro-average AUC）。

图2 基于2598例次肺癌临床症状及体征的9种基础模型与融合模型效能对比

Figure 2 Comparison of the performance of base models and fusion models for clinical symptoms and signs of 2598 cases of lung cancer

能力，降低过拟合风险<sup>[36]</sup>，为弥补以单个传统机器学习算法为主流的传统证候诊断模型之缺陷提供可能思路。本研究基于病证结合模式，以肺癌为例，使用Stacking集成算法对中医证候诊断模型进行训练并测试，初步论证了Stacking集成算法弥补传统证候诊断模型缺陷的现实意义。本研究收集了全国9家医院提供的2598例次数据，覆盖125种临床症状及体征，13种中医证候，并对SVM、KNN、RF、ExtraTrees、XGBoost、LightGBM、AdaBoost、GB及MLP共9种基础模型进行训练，将表现较优的4种基础模型（RF、ExtraTrees、MLP、SVM）进行融合，再将上述4种基础模型的预测结果堆叠并匹配证候后再次使用以上9种算法进行训练，最终得到9种融合模型。在对融合模型的评估中，训练集使用XGBoost(0.850)的准确率低于RF(0.854)，测试集使用XGBoost(0.838)的准确率低于KNN(0.855)和RF(0.840)，但XGBoost(0.012)的过拟合差异优于RF(0.014)和ExtraTrees(0.027)，而KNN(-0.002)的过拟合差异为负，即存在欠拟合风险。同时，XGBoost的micro-average AUC最高为0.996，其混淆矩阵也证明了XGBoost融合模型在证候局部和整体判别性能上的稳定性。对9种融合模型预测各证候AUC的标准差进行计算，得到XGBoost(0.003)标准差最小，提示预测精度波动性最小，对13种中医证候的预测能力最稳定。此外，所有融合模型在测试集中的准确率和micro-average AUC的平均改善率分别为117.32%和16.68%，相比基础模型均有所改善，表明Stacking集成算法

训练对中医证候诊断融合模型效能的优化作用具有普适性。本研究结果揭示了基于Stacking集成算法的融合模型在提升肺癌中医证候诊断效能方面具有显著优势，且该方法具备良好的普适性，可为同类问题的优化提供参考。

相较于肺癌证候诊断基础模型，使用Stacking集成算法构建的肺癌证候诊断融合模型虽表现出较好的优化作用，但无可避免增加了模型训练的计算成本和模型参数的复杂度。另在融合模型训练中，由于算力限制，本研究对4种基础模型使用统一权重融合训练，暂未对模型融合权重分配策略进行更细致的探究，后续研究将针对模型权重分配策略问题开展优化研究。

本研究通过Stacking集成算法整合了多种基础模型的优势，从而更好适应中医证候数据的不平衡、非线性等特点，显著提高人工智能在中医领域精准辨证的能力，为提高肺癌临床辨证准确性和效率提供了有力支持，也为肺癌中医个体化诊疗方案的自动化制定提供前瞻性保障，后续研究团队将扩大病种范围，更深层次论证Stacking集成算法的准确性、稳定性和可复制性，以期同类研究提供更丰富的方法学参考和实证基础。

参考文献

[1]沈舒文，宇文亚．中医辨证论治与标准化问题的思考[J]．中华中医药学刊，2013，31(10)：2088-2090．  
[2]赵国桢，郭诗琪，庞华鑫，等．人工智能技术在辅助中医诊疗及诊疗标准化中的应用[J]．中医杂志，2022，63(24)：2306-2310．

- [3]黄欣荣,钟平玉,马纲.人工智能与中医智能化[J]. 中医杂志, 2017, 58(24): 2076-2079, 2106.
- [4]余楷杰,袁芳君,马庆宇,等.机器学习驱动中医诊断智能化的发展现状、问题及解决路径[J]. 中国中医基础医学杂志, 2024, 30(3): 398-406.
- [5]DŽEROSKI S, ŽENKO B. Is combining classifiers with stacking better than selecting the best one? [J]. Mach Learn, 2004, 54(3): 255-273.
- [6]SUGIYAMA M. Introduction to statistical machine learning[M]. Boston: Morgan Kaufmann, 2016: 343-354.
- [7]中华医学会,中华医学学会肿瘤学分会,中华医学学会杂志社. 中华医学学会肺癌临床诊疗指南(2019版)[J]. 中华肿瘤杂志, 2020, 42(4): 257-287.
- [8]中华中医药学会. 中医临床诊疗术语 第2部分:证候[EB/OL]. (2022-03-15)[2023-12-01]. <https://www.cacm.org.cn/2022/03/15/17240>.
- [9]HUANG Y, ZHU T, ZHANG X, et al. Longitudinal MRI-based fusion novel model predicts pathological complete response in breast cancer treated with neoadjuvant chemotherapy: a multicenter, retrospective study[J]. EClinicalMedicine, 2023, 58: 101899. doi: 10.1016/j.eclinm.2023.101899.
- [10]CUI Y, ZHANG J, LI Z, et al. A CT-based deep learning radiomics nomogram for predicting the response to neoadjuvant chemotherapy in patients with locally advanced gastric cancer: a multicenter cohort study[J]. EClinicalMedicine, 2022, 46: 101348. doi: 10.1016/j.eclinm.2022.101348.
- [11]GAL MS, RUBINFELD DL. Data standardization[J]. NYUL Rev, 2019, 94:737. doi: 10.2139/ssrn.3326377.
- [12]SCHÖBER P, BOER C, SCHWARTE LA. Correlation Coefficients: Appropriate Use and Interpretation [J]. Anesth Analg, 2018, 126(5):1763-1768.
- [13]LI Z, SILLANPÄÄ MJ. Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection[J]. Theor Appl Genet, 2012, 125(3): 419-435.
- [14]徐玮斐,刘国萍,王忆勤,等.近5年中医证候诊断客观化研究述评[J]. 中医杂志, 2016, 57(5): 442-445.
- [15]宋召青,陈垚.基于支持向量机的多类分类算法综述[J]. 海军航空工程学院学报, 2015, 30(5): 442-446.
- [16]林香亮,袁瑞,孙玉秋,等.支持向量机的基本理论和研究进展[J]. 长江大学学报(自科版), 2018, 15(17): 48-53, 6.
- [17]THAKKAR HK, LIAO WW, WU CY, et al. Predicting clinically significant motor function improvement after contemporary task-oriented interventions using machine learning approaches[J]. J Neuroeng Rehabil, 2020, 17(1): 131.
- [18]杨巍,文小平,郭晶磊.以法统方结合机器学习探索中医温胆治法沿革[J]. 中国中医药信息杂志, 2020, 27(11): 96-98.
- [19]李杨,陈子彬,谢光强.一种基于ExtraTrees的差分隐私保护算法[J]. 计算机工程, 2020, 46(2): 134-140.
- [20]刘尚松.面向点击率预测的梯度提升决策树可视分析研究[D]. 成都:四川大学, 2023.
- [21]吴晖南,陈淑娇,陈展峰,等.基于LightGBM模型的糖尿病预测模型的研究[J]. 中国卫生标准管理, 2023, 14(24):64-67.
- [22]蒋晋文,刘伟光. XGBoost算法在制造业质量预测中的应用[J]. 智能计算机与应用, 2017, 7(6): 58-60.
- [23]HASTIE T, ROSSET S, ZHU J, et al. Multi-class AdaBoost[J]. Stat Interface, 2009, 2(3): 349-360.
- [24]王婧萱,王雯婧,闻亮,等.基于深度学习与多模态生理数据的阿尔茨海默病分类方法研究[J]. 医疗卫生装备, 2023, 44(11):1-8.
- [25]SWETS JA. Measuring the accuracy of diagnostic systems [J]. Science, 1988, 240(4857): 1285-1293.
- [26]MANDREKAR JN. Receiver operating characteristic curve in diagnostic test assessment[J]. J Thorac Oncol, 2010, 5(9): 1315-1316.
- [27]LOBO JM, JIMÉNEZ-VALVERDE A, REAL R. AUC: a misleading measure of the performance of predictive distribution models [J]. Global Ecol Biogeogr, 2008, 17(2): 145-151.
- [28]HEYDARIAN M, DOYLE TE, SAMAVI R. MLCM: multi-label confusion matrix [J]. IEEE Access, 2022, 10: 19083-19095. doi:10.1109/ACCESS.2022.3151048.
- [29]毋心怡. 基于 Logistic 回归与集成学习方法的代谢综合征中医证候识别与诊断研究[D]. 广州:暨南大学, 2022.
- [30]高振,董竞成. 由四诊合参到中医精准辨证论治[J]. 中华中医药杂志, 2019, 34(1): 13-17.
- [31]全小林,何莉莎,赵林华. 中医迈向精准时代的思考[J]. 中医杂志, 2016, 57(20): 1715-1718.
- [32]田甜,张培彤. 肺癌证候研究方法现状及展望[J]. 中国中医药信息杂志, 2010, 17(4): 105-108.
- [33]李建生,春柳,冯贞贞,等. 基于关联规则结合贝叶斯网络的症状间关系及其对证候诊断贡献度的研究: 以新型冠状病毒肺炎湿热蕴肺证为例[J]. 中华危重病急救医学, 2020, 32(9): 1045-1050.
- [34]付钊. 基于文本语义分块的中医病情分类问题研究[D]. 杭州:浙江大学, 2018.
- [35]罗慤婧,柴倩云,冯玉婷,等. 中医药临床研究证候

标准化诊断的研究思路与方法[J]. 中医杂志, 2023, 64(24): 2505-2510.

[36]刘洋. 基于 Stacking 融合模型的高血压风险预测研究[D]. 武汉: 华中农业大学, 2023.

## Establishment of a Traditional Chinese Medicine Syndrome Diagnostic Model Based on Stacking Ensemble Learning: Take Lung Cancer as an Example

GUO Xiaochuan<sup>1,2,3</sup>, FENG Zhenzhen<sup>1,2</sup>, LIU Wenrui<sup>1,2,3</sup>, LI Jiansheng<sup>1,2</sup>

1. Co-construction Collaborative Innovation Center for Chinese Medicine and Respiratory Diseases by Henan and Education Ministry of P. R. China, Henan University of Chinese Medicine, Zhengzhou, 450046; 2. The First Affiliated Hospital of Henan University of Chinese Medicine; 3. The First Clinical Medical College, Henan University of Chinese Medicine

**ABSTRACT Objective** To explore the method of optimizing the performance of traditional Chinese medicine (TCM) syndrome diagnostic models using Stacking ensemble learning. **Methods** Taking the construction of TCM syndrome diagnostic model for lung cancer as an example, 2598 cases of clinical symptoms and signs from lung cancer patients in 9 hospitals were used as independent variables (i. e., feature variables), TCM syndrome information as dependent variables, and the clinical data were divided into training set and testing set in 8:2 ratio according to random number table method using Python 3.7 software. The stable features of TCM syndrome of lung cancer were screened using chi-square test, Spearman's correlation test, and Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression analysis; nine machine learning algorithms are trained, including support vector machines (SVMs), k-nearest neighbors (KNN) algorithm, Random Forest (RF), Extremely Randomized Trees, Extreme Gradient Boosting (XGBoost), Lightweight Gradient Boosting (LightGBM), Adaptive Boosting (AdaBoost), Gradient Boosting (GB) and the multi-layer perceptron (MLP), to obtain 9 basic models. Four models with better performance were screened out from the above basic models and fused to form a fusion model by using the Stacking ensemble learning, and the fusion model was trained twice by the above nine machine learning algorithms and evaluated by accuracy rate, micro-average ROC curves, area under the curve (AUC), and confusion matrix metrics, to screen the optimal diagnostic model. **Results** After data processing, 79 stable features and 13 TCM syndromes were obtained. In the basic model training, the comprehensive performance of RF, ExtraTrees, MLP and SVM basic models were better, so the predicted distributions of the syndromes of these four models were used as the secondary training data, and nine fusion models were obtained based on the Stacking ensemble learning (SVM, KNN, RF, ExtraTree, XGBoost, LightGBM, GB, AdaBoost, MLP). Among them, the XGBoost fusion model performed the best, with an accuracy of 0.850 and 0.838 in the training set and test set, respectively, an overfitting difference of 0.012, and an area under the micro-average ROC curve of 0.996. All fusion models showed an improvement in accuracy and area under the micro-average ROC curve compared with the base model in the test set. **Conclusion** Taking the TCM syndrome information of lung cancer as an example, the XGBoost fusion model has significant advantages in improving the diagnostic performance of TCM syndrome information of lung cancer through Stacking ensemble learning. It can be seen that the advantages of Stacking ensemble learning to integrate multiple models and effectively improve the diagnostic efficiency of TCM diagnostic models, which provided a methodological reference for similar studies.

**Keywords** diagnostic model of traditional Chinese medicine syndrome; lung cancer; syndrome; machine learning; Stacking ensemble learning

(收稿日期: 2024-02-12; 修回日期: 2024-05-04)

[编辑: 焦 爽]